Initial Observations on Query Based Sampling in Distributed CLIR

Xiao Mang Shou, Mark Sanderson
Department of Information Studies
University of Sheffield
Western Bank, Sheffield S10 2TN, UK
[x.m.shou, m.sanderson]@shef.ac.uk

Abstract

Cross Language Information Retrieval (CLIR) enables people to search information written in different languages from their query languages. Information can be retrieved either from a single cross lingual collection or from a variety of distributed cross lingual sources. This paper presents initial results exploring the effectiveness of distributed CLIR using query-based sampling techniques, which to the best of our knowledge has not been investigated before. In distributed retrieval with multiple databases, query-based sampling provides a simple and effective way for acquiring accurate resource descriptions which helps to select which databases to search. Observations from our initial experiments show that the negative impact of query-based sampling on cross language search may not be as great as it is on monolingual retrieval.

1 Introduction

Cross Language Information Retrieval (CLIR) is the process of retrieving documents written in a language(s) different from the language of the query. In recent years much CLIR research was undertaken in the academic communities via the academic evaluation forums like CLEF, NTCIR and TREC and a number of application areas have been developed over the years. CLIR is basically a combination of machine translation and traditional monolingual IR and four approaches commonly used for translation include [Gollins, 2000]: (1) a controlled vocabulary, (2) machine translation, (3) bilingual parallel corpora, (4) bilingual dictionaries, or more recently a combination of all approaches. When doing translation, one can either translate the query into the target language (query translation, QT), translate search documents into the query language (document translation, DT), or translate both queries and documents into a common language [Oard, 1997]. So far, query translation, which transforms a user's query into the language of the documents, is the dominant approach because this can be made to work successfully with simple translation methods and does not require the overhead of translating collection documents which is often computationally expensive. With the right approach, CLIR systems are able to achieve retrieval effectiveness that is only marginally degraded from the effectiveness achieved had the query been manually translated [Ballesteros & Croft, 1998].

Current CLIR research focuses on improving retrieval effectiveness under monolingual, bilingual or multilingual conditions. However, how to process multilingual information in a distributed environment has not yet been sufficiently explored. In distributed retrieval with multiple multilingual resources (referred to here as databases), the common approach is to translate queries into the resource language for retrieval and then results from individual collections are merged into a single list. Using this method, similar to monolingual distributed retrieval, when there are a large number of databases, it can be difficult to choose which databases to search. This situation exacerbates in a multilingual environment. Obtaining cross language resource descriptions of each database automatically and efficiently becomes necessary.

Query-based sampling (QBS) [Callan and Connell, 2001] is a technique used for acquiring resource descriptions of databases by running queries on the databases examining text of the documents returned, seeking new queries from the text and using the text to build a uni-gram language model of the database content. Empirically, results demonstrated that sampling 300-500 documents from each database appears to be effective for resource description across different rage of database sizes. This method is particularly useful in distributed retrieval to decide which databases to search according to a given query. To the best of our knowledge, there has been no investigation of QBS and CLIR in the past. Therefore, in our experiments, we tested QBS with query translation, document translation and both query and document translation together for distributed retrieval using the CLEF2003 Italian collection. QBS Results were compared with original distributed monolingual retrieval results. What we report here are a series of observations based on our initial experiments.

This paper divides into the following sections: section 2 describes our experimental set up, section 3 presents our results and compares these results with original monolingual baseline, section 4 list some directions for future work and section 5 summaries our findings.

2 Experimental Setup

2.1 The CLEF2003 Test Collection

The test data set we use is CLEF (http://clef.isti.cnr.it/) 2003's Italian collection (157,558 documents, average

document length 214, overall size 370MB) and the 60 query topics (141- 200) with title, description and narrator fields. The same 60 query topics are available in eight languages including Italian and English by human translation.

For additional comparison, CLEF2003's English collection was tested for query translation using both original English topics and machine translated Italian to English topics.

2.2 Distributed Retrieval and Query-Based Sampling

In our experiments, we tested cross language retrieval using distributed retrieval methods where the original Italian collection was divided into 30 sub databases by document order in collection with each of them containing around 5,252 documents. Distributed retrieval was performed based on the 30 databases using both complete and sampled resource descriptions. Resource descriptions store information about what each database contains. A complete resource description is generated using full collection index whereas sampled resource description (sometimes called learned resource description) is generated by QBS. In our experiments, for each of the 30 sub databases, 300 documents were selected to obtain the sampled resource description.

In parallel, the same set of retrievals were run for the translated Italian to English collection and CLEF2003 English collection as well for further comparison.

2.3 Translation Resources and Retrieval System

The machine translation tool used in our experiments was Systran Professional Premium 5.0's MultiTranslate Utility (http://www.translation.net/systran_professional.html) which can translate multiple files in batches. The whole Italian collection was translated into English and this process took about one month running on a "standard desktop PC". Using the same tool, original Italian query topics were translated into English and English topics were translated into Italian.

The retrieval tool we used for Italian and English retrieval was Lemur3.1 (http://www.lemurproject.org/). Lemur supports distributed retrieval providing functions to rank databases by their resource descriptions and merge their distributed search results using the CORI algorithm [Callan, 2000]. The default setting for Lemur in our experiments was to retrieve the top 30 ranked documents from the top 10 ranked databases.

Since CLEF data format was not compatible with Lemur and Systran formats, conversion of the query and document format was necessary.

2.4 Cross-Language Retrieval

Given the Italian collection and queries and their translated English versions, we compare query versus document translation alone as well as applying both query and document translation. Before the experiment, the data collection and queries were processed. We first translated the Italian text and queries into English and English queries into Italian using the MT system. Next, stopwords were removed using stopword lists provided by the Snowball stemmer (http://snowball.tartarus.org). We then applied stemming using Snowball and removed diacritics in Italian using the UNIX recode tool. To perform this, we recoded the character set from latin1 to HTML and then replaced the HTML characters by their original ASCII characters. Finally, all characters were converted to lower case. After the process, all collection texts were split into sub databases to be indexed and retrieved by Lemur.

The following experiments were performed with each of them applied to distributed retrieval using both complete and sampled resource descriptions:

1. Retrieval using the original Italian collection and topics. This will be used as a baseline in comparison (monolingual).
2. Query translation with original Italian collection and English topics translated to Italian (QT).
3. Document translation with the Italian collection translated to English and original English topics (DT).
4. Both query and document transition with the Italian collection translated to English and the Italian topics translated to English (QT+DT).

The English collection was tested under monolingual and query translation configurations. Precision at rank 5, 10, 15, 20 and 30 (P5, P10, P15, P20, P30) was used to measure retrieval effectiveness. In addition, recall and the number of queries with any relevant documents retrieved (rel_q) was also computed.

3 Results

Results are listed in Tables 1, 2 and 3: the tables show results covering a number of configurations of the collection and topics. Each table collates the results for a particular form of collection: in Table 1, the native Italian collection; table 2, native English; and in table 3, Italian collection translated to English. In each table two forms of query are shown and within each query the search on the full (columns 1&3) and sampled resource descriptions (columns 2&4) are compared. Since we did not have time to translate the English collection to Italian, results for that configuration are not shown.

	Italian collection, Italian topics (monolingual)		Italian collection, English -> Italian topics (QT)	
	1	2	3	4
	Complete resource description	Sampled resource description	Complete resource description	Sampled resource description
P5	0.3133	0.3067 (-2.1%)	0.2667	0.2533 (-5.0%)
P10	0.2483	0.2333 (-6.0%)	0.2033	0.2067 (+1.7%)
P15	0.2111	0.2033 (-3.7%)	0.1711	0.1744 (+2.0%)
P20	0.1867	0.1808 (-3.2%)	0.1467	0.1467 (0.0%)
P30	0.1511	0.1406 (-7.0%)	0.1128	0.1183 (+4.9%)
rel_q	43/51	41/51 (-4.7%)	39/51	40/51 (+2.5%)
recall	272/809	253/809 (-7.0%)	203/809	213/809 (+4.9%)

Table 1. Italian monolingual and QT using complete and sampled resource descriptions

	English collection, English topics (monolingual)		English collection, Italian -> English topics (QT)	
	1	2	3	4
	Complete resource description	Sampled resource description	Complete resource description	Sampled resource description
P5	0.4100	0.3367 (-10.6%)	0.2933	0.2500 (-14.8%)
P10	0.3100	0.2817 (-9.1%)	0.2250	0.2050 (-8.9%)
P15	0.2633	0.2456 (-6.7%)	0.1989	0.1878 (-5.6%)
P20	0.2400	0.2200 (-8.3%)	0.1817	0.1675 (-7.8%)
P30	0.2028	0.1806 (-11.0%)	0.1572	0.1411 (-10.2%)
rel_q	47/54	44/54 (-6.4%)	42/54	39/54 (-7.1%)
recall	365/1006	325/1006 (-11.0%)	283/1006	254/1006 (-10.2%)

Table 2. English monolingual and QT using complete and sampled resource description

	Italian -> English collection, English topics (DT)		Italian -> English collection, Italian -> English topics (QT+DT)	
	1	2	3	4
	Complete resource description	Sampled resource description	Complete resource description	Sampled resource description
P5	0.2133	0.1867 (-12.5%)	0.3133	0.2467 (-21.3%)
P10	0.2017	0.1800 (-10.8%)	0.2683	0.2133 (-20.1%)
P15	0.1833	0.1489 (-18.8%)	0.2322	0.1911 (-17.7%)
P20	0.1625	0.1342 (-17.4%)	0.2008	0.1667 (-17.0%)
P30	0.1283	0.1111 (-13.4%)	0.1611	0.1322 (-17.9%)
rel_q	39/51	37/51 (-5.1%)	46/51	41/51 (-10.7%)
recall	231/809	200/809 (-9.7%)	290/809	238/809 (-17.9%)

Table 3. Italian DT and QT+DT using complete and sampled resource description

As can be seen across all the tables, with one exception (in Table 1), query-based sampling reduces effectiveness compared to retrieval based on the full resource description. The one exception to this is under the QT condition where use of the sampled resource description resulted in improved effectiveness (average of the % difference in column 4 of Table 1 is +0.72%). In addition, more relevant documents were retrieved and more queries with at least some relevant documents retrieved were found than with the full resource description. However, the improvement was not significant. In general, as would be expected, QBS reduced retrieval effectiveness. The reductions observed in Tables 1 & 2 are in-line with reductions reported from the original QBS paper [Callan and Connell, 2001].

The reductions in Table 3 are larger. Here collection translation is being used, with the Italian collection being translated into English. Errors will be made in the translation process and whether those errors are somehow causing problems in the resource description process of QBS, is an area of investigation to be examined in the future.

3.1 Using translation to enhance monolingual search?

Separate from QBS, we report one other result. Comparing column 3 of Table 3 with column 1 of Table 1, both configurations are the same taking Italian queries retrieving on the same CLEF Italian collection. The results in the column of Table 3 however shows monolingual retrieval after both queries and documents are translated into another language, in this case English. Performing QT and DT together resulted in a 6.4% precision improvement on average over original monolingual retrieval with no translation. This gain is consistent after rank 10, but not significant. Even under QBS condition, comparing results in column 2 in Table 1 with column 4 in Table 3, performing QT and DT together resulted in 9.6% precision drop on average over monolingual retrieval which is still comparably effective.

Our observation that translating both collection and queries from Italian to English together outperforms its originally monolingual baseline is to the best of our knowledge new. Franz and McCarley's tried improving monolingual retrieval by query and document translation (as done here) working on an English test collection, who's documents and queries were translated into French [Franz and McCarley, 1999]. With their experiments, however, they reported a 10-20% drop in effectiveness compared to monolingual baseline.

Whether our observation is an outlier case or an indication of a general trend will be the subject of further work.

4 Future Work

In our experiments, we only managed to translate the CLEF2003 Italian collection and topics into English and English topics to Italian. Due to the restriction of collection, the query-based sampling size is approximately 5% of each database and may be impractical for larger collections. In order to better understand the importance, significance and durability of our observations, we will extend our work to larger collection size and other language translations. Further future work includes translating other languages such as German into English and observes whether the specialty of compound words in German will result in different performance in CLIR. Since document translation are time consuming, other translation resources such as dictionary look up, parallel texts based translation [Chen and Gey, 2003] or statistical translation models built by GIZA++ (http://www.fjoch.com/GIZA++.html) can be applied for future experiments as well. We also will translate English collection to Italian and run the same set of experiments on it to test the bilingual distributed CLIR and to better understand the drops in effectiveness when using query-based sampling and document translation.

5 Conclusion

In this paper we have presented experiments incorporating machine translation of both the queries and documents into an IR engine for distributed CLIR using Systran Professional and CLEF2003 Italian collection. Distributed CLIR was run using a complete resource description of all collections or using sampled resource collection by query-based sampling technique. The obtained results were compared with the results from original monolingual retrieval and results from applying query and document translation. In Italian distributed CLIR, we observed that QT using QBS sampled resource description performed better than DT at the same condition, and its performance

was comparably effective as using complete resource description. Furthermore, we also found that when translating both queries and documents together, distributed cross language retrieval is almost as good as monolingual retrieval either using complete resource description or QBS sampled resource description.

Acknowledgments

The work described in this paper was funded as part of the EU 6th Framework project, BRICKS - Building Resources for Integrated Cultural Knowledge Services. Further information on the project can be found at http://www.brickscommunity.org/.

References

[Ballesteros and Croft, 1998] Lisa Ballesteros and Bruce Croft. Resolving Ambiguity for Cross Language Retrieval. In *Proceedings of the 21st International Conference on Research and Devlelopment in Information Retrieval*, p64-71.

[Franz and McCarley, 1999] Martin Franz and J. Scott McCarley. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, p295-296.

[Gollins, 2000] Timothy John Collins. Dictionary Based Transitive Cross-Language Information Retrieval using Lexical Triangulation. *Masters Dissertation.* Department of Information Studies, University of Sheffield.

[Oard, 1997] Doug Oard. Serving Users in Many Languages. D-Lib.

[Callan, 2000] Jamie Callan. Distributed Information Retrieval. W. B. Croft Editor, Kluwer Acedemic Publishers 2000, *Advances in Information Retrieval*, p127-150

[Callan and Connell, 2001] Jamie Callan and Margaret Connell. Query-Based sampling of Text Databases. In *ACM Transactions on Information Systems (TOIS)*, Volume 19 , Issue 2 (April 2001), p97-130

[Chen and Gey, 2003] Aitao Chen, Fredric C. Gey. Combining Query Translation and Document Translation. in Cross-Language Retrieval. In *CLEF 2003*, p108-121